TP-OOD Report

Baseline Training, OOD Detection, Neural Collapse Analysis, and NECO

Yuheng ZHANG & Qizheng WANG
ENSTA Paris

February 19, 2026

Abstract

This report presents the work carried out for the out-of-distribution (OOD) detection practical based on the OpenOOD framework. We first trained a ResNet-18 classifier on CIFAR-100 to obtain a reference in-distribution model. Several OOD scoring methods were then evaluated using the unified OpenOOD evaluation pipeline. In parallel, we investigated Neural Collapse phenomena along the training trajectory through a set of geometric diagnostics. Finally, we implemented the NECO (Neural Collapse Inspired OOD Detection) method and compared its performance with standard approaches. The report focuses on the main implementation choices and qualitative observations rather than exhaustive hyperparameter tuning.

Experimental Setup

All experiments were conducted using the OpenOOD framework, which provides a modular interface for model training and OOD evaluation. CIFAR-100 was used as the in-distribution (ID) dataset throughout the study. To assess OOD behavior, we relied on a combination of near-OOD datasets (CIFAR-10 and TinyImageNet) and far-OOD datasets (MNIST, SVHN, Texture, and Places365). This separation allows us to distinguish between distribution shifts that are visually similar to CIFAR-100 and those that differ more substantially.

The backbone model considered in this work is ResNet-18 adapted to 32×32 inputs. Feature vectors extracted from the penultimate layer play a central role in both Neural Collapse analysis and the NECO method. OOD performance is quantified using standard metrics, including AUROC, FPR@95, and AUPR.

1 Baseline Training on CIFAR-100

The first stage of the project consisted of training a standard classifier on CIFAR-100. OpenOOD's baseline training pipeline was used with minimal modifications. Checkpoints were saved periodically to enable later analysis of training dynamics. The training process converged smoothly, producing a final model with an in-distribution test accuracy of approximately 77.42%.

The evolution of the loss and validation accuracy followed the expected pattern for this architecture and dataset. No unusual instabilities were observed, which suggests that the resulting model is a suitable basis for subsequent OOD experiments.

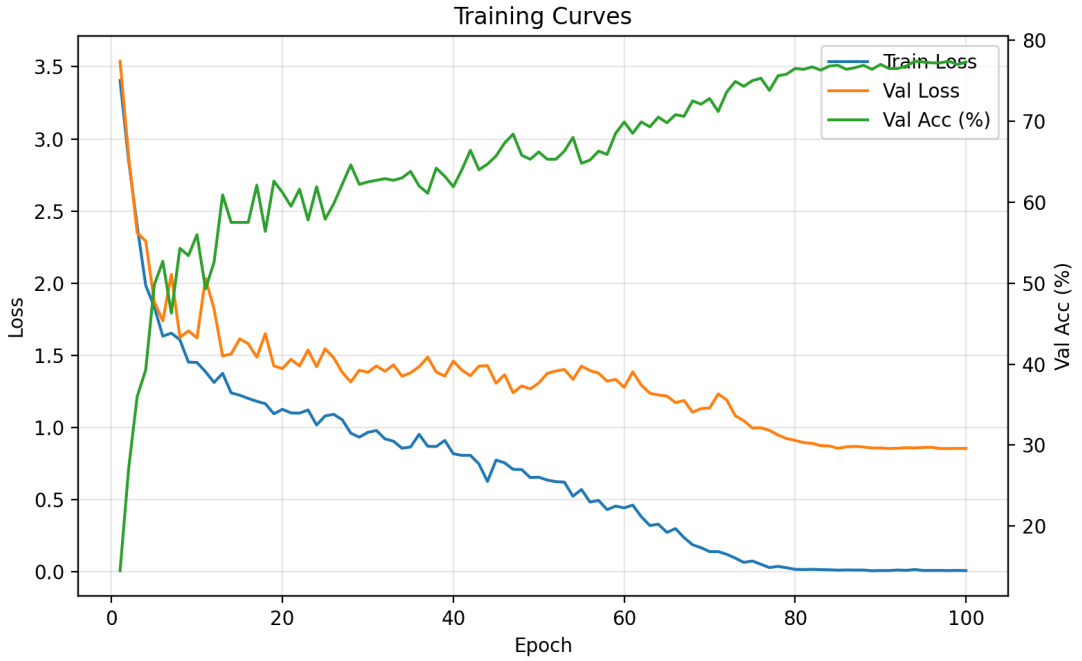

Figure 1: Training dynamics of ResNet-18 on CIFAR-100. We show the learning curves for loss and validation accuracy during training.

2 OOD Detection with Standard Scores

With the baseline model fixed, we evaluated several widely used OOD scoring methods available in OpenOOD. These include Maximum Softmax Probability (MSP), MaxLogit, Mahalanobis distance, Energy-based scores, and ViM. The unified evaluation script provided by OpenOOD greatly simplifies this step by handling dataset loading, inference, and metric computation in a consistent manner.

Qualitatively, the results align with common observations in the OOD literature. Near-OOD datasets such as CIFAR-10 tend to be more challenging, as reflected by lower AUROC values and higher FPR@95. In contrast, far-OOD datasets typically exhibit clearer separability from the ID distribution. These differences are useful for interpreting the behavior of more specialized methods such as NECO.

Method	Near-OOD AUROC ↑	Far-OOD AUROC ↑
MSP	80.52	77.89
MaxLogit (MLS)	81.34	79.10
Mahalanobis (MDS)	57.97	71.47
Energy (EBO)	81.17	78.94
ViM	75.22	81.81

Table 1: Comparison of standard OOD scoring methods on CIFAR-100 (ID). We report mean AUROC on Near-OOD (CIFAR-10 + TinyImageNet) and Far-OOD (MNIST + SVHN + Texture + Places365), as provided by OpenOOD.

3 Neural Collapse Diagnostics

A significant part of this project was devoted to studying Neural Collapse properties along the training trajectory. Neural Collapse describes a set of geometric regularities that often emerge near the end of training, including reduced within-class variability, structured class means, and alignment between classifier weights and features.

To investigate these effects, we computed a collection of metrics (NC1–NC4) using intermediate checkpoints saved during training. These metrics capture complementary aspects of feature geometry, such as the relative magnitude of within-class scatter, the coherence of class means, the alignment between classifier weights and class means, and the agreement between nearest class center predictions and network outputs.

Across epochs, the general trends are consistent with the Neural Collapse hypothesis. For instance, the metrics indicate that the representation becomes more structured over training and the classifier aligns better with feature statistics.

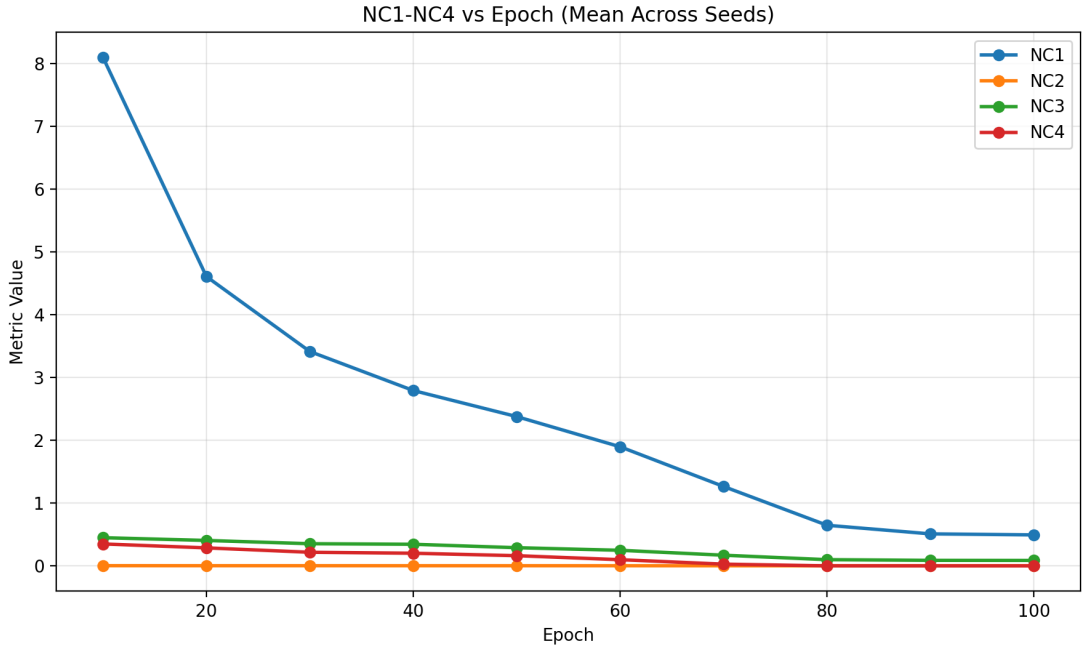

Figure 2: Evolution of Neural Collapse metrics (NC1–NC4) during training.

To make the alignment aspect more explicit, we also report the cosine similarity between classifier weights and class mean features. A rising trend here is consistent with the NC3 phenomenon (weight/mean alignment) in the terminal phase.

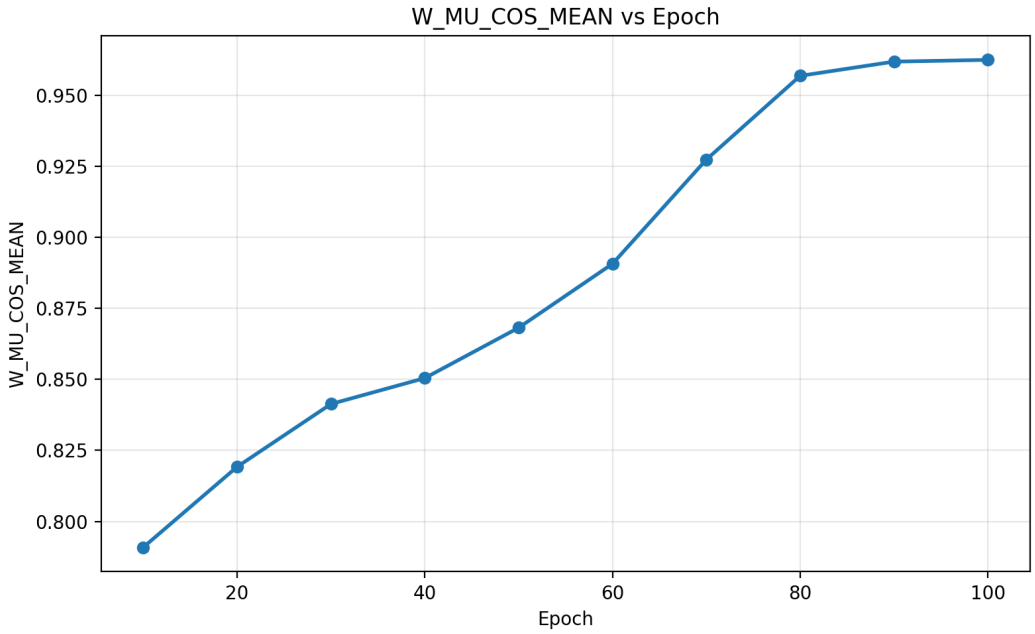

Figure 3: Cosine similarity between classifier weights and class mean features across epochs, illustrating the progressive alignment predicted by Neural Collapse.

4 NC5 and ID/OOD Geometry

Beyond the classical Neural Collapse metrics, we also examined NC5, which measures the alignment between OOD feature means and ID class means. This metric offers a simple perspective on how OOD samples are positioned relative to the learned ID structure. Lower NC5 values may indicate a greater degree of orthogonality between ID and OOD representations.

Although NC5 is not directly used for classification or detection, it provides an intuitive diagnostic tool. In our experiments, NC5 values varied across OOD datasets in a manner that broadly reflects their visual similarity to CIFAR-100.

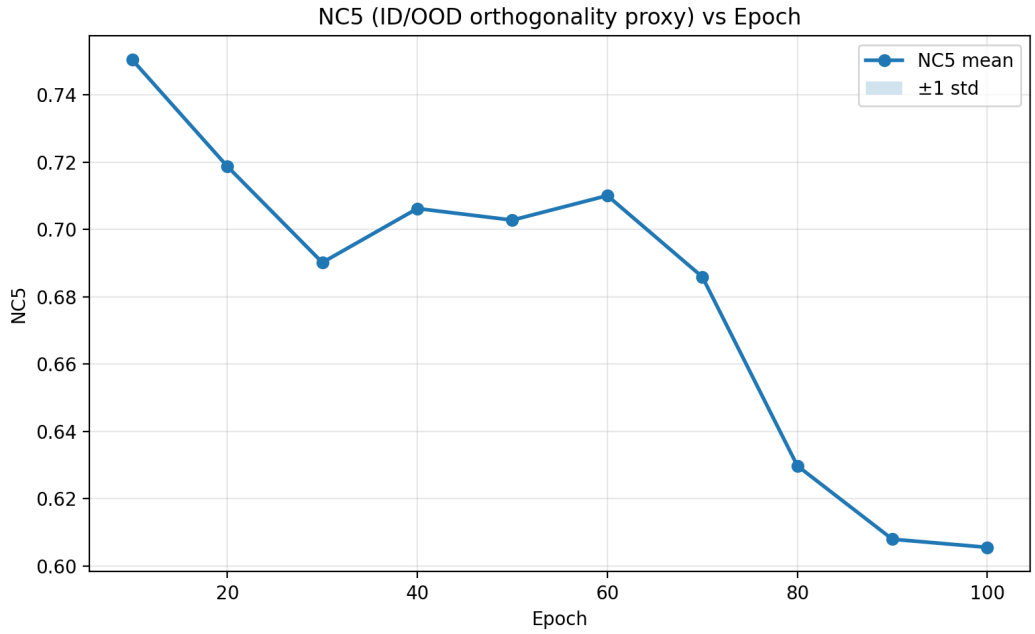

Figure 4: NC5 across epochs (aggregated over the selected OOD split). Lower values suggest that the mean OOD representation becomes more "orthogonal" to ID class means.

5 NECO: Neural Collapse Inspired OOD Detection

Motivated by Neural Collapse observations, we implemented the NECO method as a custom OpenOOD postprocessor. The main idea behind NECO is that, near convergence, ID features are often well described by a low-dimensional subspace. By projecting features onto the principal components obtained from ID training data, we can derive a simple OOD score based on projection strength.

More precisely, penultimate-layer features are optionally standardized and used to fit a PCA model. The NECO score is defined as the ratio between the norm of the projected feature and the norm of the full feature vector. This quantity serves as a proxy for how strongly a sample aligns with the dominant ID feature structure.

As a sanity check and for interpretability, we visualize the penultimate-layer features with a 2D PCA projection. Even though PCA is a strong dimensionality reduction, it already provides an intuitive separation between ID and far-OOD samples.

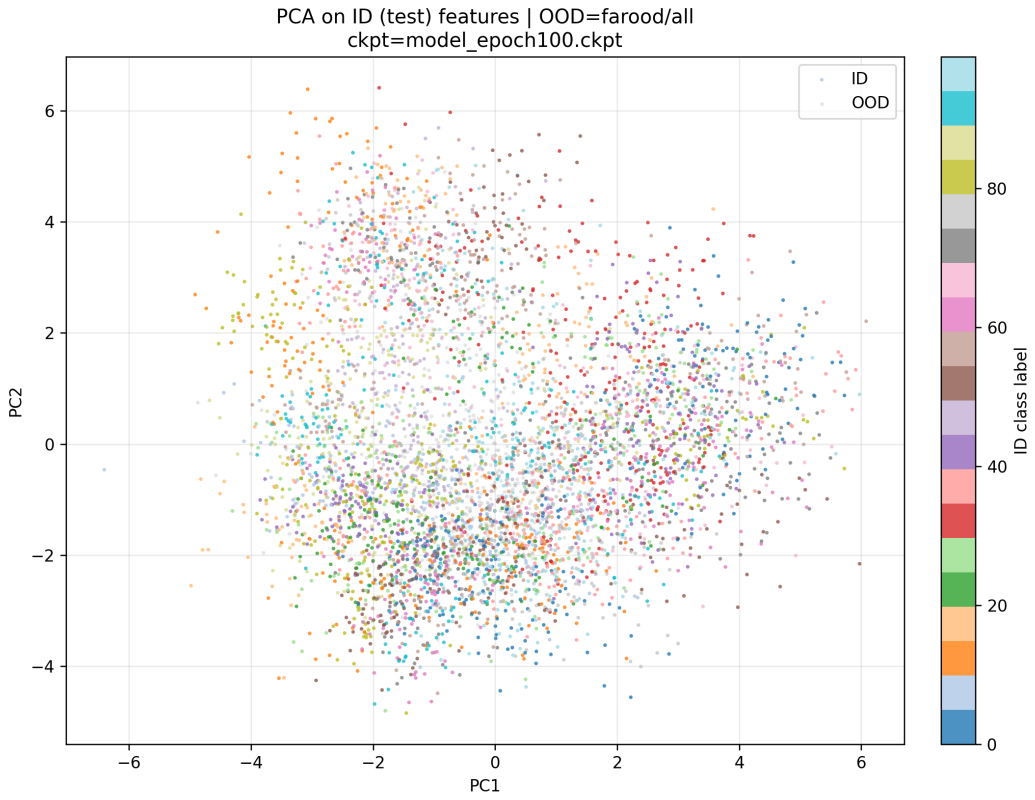

Figure 5: PCA visualization of penultimate-layer features for ID test samples and far-OOD samples. This motivates NECO, which uses alignment with the principal ID subspace as an OOD signal.

The method was integrated into OpenOOD's evaluation pipeline, allowing direct comparison with existing scores. The resulting metrics exhibit the expected pattern: far-OOD datasets are detected more reliably than near-OOD datasets. The numerical values are within a reasonable range given the simplicity of the approach and the difficulty of near-OOD detection.

Dataset	FPR@95	AUROC
Near-OOD (mean)	72.49	72.27
Far-OOD (mean)	53.09	82.51

Table 2: Summary NECO results (single run).

6 Conclusion

This practical provided a structured introduction to OOD detection within a modern evaluation framework. Starting from a standard CIFAR-100 classifier, we explored both established OOD scoring techniques and geometric diagnostics derived from Neural Collapse theory. The NECO method, while conceptually simple, illustrates how representation geometry can be leveraged for OOD detection.